

Claremont Colleges

Scholarship @ Claremont

Library Staff Publications and Research

Library Publications

11-4-2016

Critical Collection Analysis: Using DH Tools to Contextualize Historical Collecting Patterns within a Political Framework

Lydia Bello

Claremont University Consortium

Nina Clements

Claremont University Consortium

Madelynn Dickerson

Claremont University Consortium

Margaret Hogarth

Claremont University Consortium

Follow this and additional works at: https://scholarship.claremont.edu/library_staff



Part of the [Cataloging and Metadata Commons](#), [Collection Development and Management Commons](#), and the [Digital Humanities Commons](#)

Recommended Citation

Bello, Lydia; Clements, Nina; Dickerson, Madelynn; and Hogarth, Margaret, "Critical Collection Analysis: Using DH Tools to Contextualize Historical Collecting Patterns within a Political Framework" (2016). *Library Staff Publications and Research*. 52.

https://scholarship.claremont.edu/library_staff/52

This Presentation is brought to you for free and open access by the Library Publications at Scholarship @ Claremont. It has been accepted for inclusion in Library Staff Publications and Research by an authorized administrator of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

Critical Collection Analysis

Using DH Tools to Contextualize Historical Collecting Patterns within a Political Framework

Lydia Bello
Nina Clements
Madelynn Dickerson
Margaret Hogarth
The Claremont Colleges Library

The Charleston Conference
November 4, 2016

CLAREMONT COLLEGES
LIBRARY • VITeL

These slides were shown at the Charleston Conference on Friday, November 4th, 2016. The presentation showcases a work in progress, and is an overview of our learning process as much as it is about the analysis we worked on.

Our critical collection analysis project began as part of an in-house professional development course for librarians in digital humanities. We spent 6 weeks learning about DH tools and during a summer “Maker Week” program, were charged with creating hypothetical projects that used DH methods to answer library related questions.

Our team includes: Lydia Bello (STEM Team Library), Nina Clements (Social Sciences Team Librarian), Madelynn Dickerson (Information Resources Coordinator), and Margaret Hogarth (Information Resources Acquisitions Team Leader). The Claremont Colleges’ digital scholarship coordinator, Ashley Sanders, developed and facilitated the DH course and “Maker Week” event, and supported our team as we worked on this project.

We were interested to find out if we could use digital humanities for collection analysis, if we could trace a topic in our collection over a finite period of time, and if we could trace a concept that crosses many call numbers. Ultimately, we thought to create a process and/or tool that other librarians could use in a meaningful way.

Overview

- DH at Claremont
- Our project
- DH and Traditional Analysis Tools
- Future applications

CLAREMONT COLLEGES
LIBRARY • VITeL

Our presentation will cover a very high level overview of DH at Claremont, an overview of our project and some specifics about DH tools that we used for our project, including Voyant Tools, TimelineJS, and Tableau. At the conclusion, we have some preliminary ideas about the possibilities for DH in the library and approaches for applying these methods to internal projects.

About Claremont

- 5 undergraduate colleges
 - Pomona College, Scripps College, Claremont McKenna College, Harvey Mudd College, Pitzer College
- 2 graduate institutions
 - Claremont Graduate University, Keck Graduate Institute
- 1 library

CLAREMONT COLLEGES
LIBRARY • **VITeL**

The Claremont Colleges Library is one library serving five undergraduate colleges and two graduate institutions with approximately 7,000 FTE. The emphasis is on liberal arts education. Each of the colleges is unique with its own characteristics, but occasionally overlapping departments and research interests. It is a perpetually complicated, and highly stimulating environment to work in and presents many challenges for the single library. Digital Humanities finds a natural home in the library as a central point for all colleges to come together for interdisciplinary research support.

DH 101 at Claremont



- 5C Mellon-funded grant to strengthen and expand DH teaching and scholarship at the colleges
- Inspired a 6 week course for librarians
- Covered definitions, data visualization, spatial & temporal pattern finding, network analysis, topic modeling

CLAREMONT COLLEGES
LIBRARY • **VITeL**

With a growing presence across the campuses, and digital humanities has become a strategic priority at the Claremont Colleges Library. Cementing this is a consortium-wide Mellon-funded grant to promote and expand DH teaching across the colleges. A new “DH Studio” is physically housed in the library. We also built the “Digital Tool Shed” this year, which opened in August 2016. The digital tool shed provides collaborative work space, a data visualization wall, 6 work stations with higher computing power and geo-spatial and other software. In addition, students can check out equipment from the library’s emerging technologies program, including Google Glass, Oculus Rift, and DSLR cameras.

A DH course for libraries was developed and taught by Ashley Sanders, PhD, our digital scholarship coordinator. Librarians from across divisions came together to learn about tools and theories in a range of DH topics, including data visualization, spatial and temporal pattern finding, network analysis, and topic modeling.

DH Course Objectives

- Develop our own definition of DH in a liberal arts context
- Know *when* to incorporate digital tools and methodologies and determine *why*
- Critically examine the strengths and weaknesses of existing tools and DH projects
- Identify sources to find and learn new digital tools and skills

CLAREMONT COLLEGES
LIBRARY • **VITeL**

The goal of the course for libraries was to build fluency in DH for effective triaging of DH related reference questions and engagement in DH-related research support. Participants hoped to be able to serve as first-tier consultants on DH projects and be able to facilitate connections to DH services in the library, and people like Ashley Sanders, our digital scholarship coordinator, for additional support.

Defining DH

- New mode of scholarship focused on collaborative, transdisciplinary, and computationally engaged research
- Known for using tools historically employed in STEM and social sciences disciplines for new scholarship
- Methods could text markup and analysis, data visualization, data mining, GIS and mapping, etc

CLAREMONT COLLEGES
LIBRARY • VITeL

There are a range of definitions for “digital humanities” and not everyone agrees. Additionally, some argue that the term “digital scholarship” is more appropriate and inclusive. Some helpful introductory definitions include the idea that DH uses computational methods and tools to answer questions and provide new avenues into humanities research and related disciplines.

DH Tools, per Matthew Milner, are any tools such as applications or software, that help gather, process, or present your research. This is a very loose definition. This definition arguably would include PowerPoint, and realistically that’s not what anyone means when they say “DH Tool,” but this example demonstrates the way these terms are in flux.

Sources for definitions and to learn more:

DH LibGuide at the Claremont Colleges Library (Ashley Sanders)
<http://libguides.libraries.claremont.edu/dh>

A Short Guide to DH (Jeffrey Schnapp)

http://jeffreyschnapp.com/wp-content/uploads/2013/01/D_HShortGuide_Page_2.jpg (Schnapp)

DH Team Projects

- Collection Analysis Team (that's us)
- Wi-Fi Analysis Team
- Fan Team
- Team Apocalypse!



CLAREMONT COLLEGES
LIBRARY • VITeL

During the in-house course, teams of libraries focused on 4 different types of projects. While our group worked on collection analysis, other teams explored data visualization using Wi-Fi data, while others used learned about network analysis by finding the relationships between theater-goers as printed on a 17th century fan and published (a work in progress) using Scalar. Another team created an interactive timeline and visual mapping of events influencing the development of 1 Enoch and Apocalyptic Literature in the Hebrew Bible.

Our Research Questions

What trends can we identify in Claremont Colleges Library collecting practices of print materials related to terrorism?

- Has print material collecting in terror studies changed over time?
- Do collecting patterns correlate to transnationally significant terror events?
- What can we learn about the books we are collecting?

CLAREMONT COLLEGES
LIBRARY • VITeL

As a part of that workshop we were discussing how libraries could use our data in these tools, and developed a proposal.

We selected terrorism as a concept because of specific events that have impacted the scholarly and national conversation. We then wanted to trace the responsiveness of our print collection to that national conversation and events.

Process Questions




- What DH tools best support this type of analysis?
- How best can we visualize our data and results?

CLAREMONT COLLEGES
LIBRARY • VITeL

As a part of our project, we had to make some decisions about what tools to use.

We ultimately decided on using Excel, Voyant-Tools, and a python script in order to scrape numbers, and Voyant Tools, TimelineJS, and Tableau to visually analyze data.

DH Tools We Experimented With

	<p>Voyant Tools https://voyant-tools.org/</p>
	<p>Timeline JS https://timeline.knightlab.com/</p>
	<p>Tableau Public http://www.tableau.com/</p>

CLAREMONT COLLEGES
LIBRARY • VITeL

All of our main tools were “out of the box” and ready to use, and many were open source.

Voyant Tools:

- Web based text reading and analysis environment
- Developed by Stefan Sinclair and Geoffrey Rockwell
- Allows user to find word frequency and context of words.

TimelineJS

- Create very visual, interactive timeline with media
- Very customizable, and was originally designed for journalism

Tableau Public

- Data visualization software, not open source.
- Tableau Public is the free version.

Complete List of Tools Used

- Innovative Millennium ILS
- Microsoft Excel
- Microsoft Access
- Voyant Tools (free online)
- Timeline JS (free online)
- Tableau Public
- Anaconda with Python Scripts by John Laudun (free online via GitHub)
- Notepad ++ (free online)

We use all free products when possible, except Microsoft products and Innovative Millennium.

Many of these tools are open source. Towards the end of the project we experimented with some more advanced tools.

Data Collection

- Innovative Millennium ILS
- Fiscal Years 1995 - 2015
- Item Records Created search
- Reported bibliographic information

Store Record Type:

Range

Term	Operator	Type	Field	Condition	Value A	Value B
1		ITEM	CREATED	between	07-01-1994	06-30-1995
2	AND	BIBLIOGRA...	BIBLVL	equal to	m	
3	AND	BIBLIOGRA...	MATTYPE	equal to	a	
4	OR	BIBLIOGRA...	MATTYPE	equal to	z	

ITEM CREATED between "07-01-1994" and "06-30-1995" AND BIBLIOGRAPHIC BIBLVL equal to "m" AND (BIBLIOGRAPHIC MATTYPE equal to "a" OR BIBLIOGRAPHIC MATTYPE equal to "z")

CLAREMONT COLLEGES
LIBRARY • VITeL

First part and arguably the most important part of the project was gathering the data. Initially we searched Innovative Millennium for order paid date by fiscal year with bibliographic value limited to books and material type limited to language materials or to e-books. We downloaded item record reports from Innovative Millennium. We searched for item records created by fiscal year, with bibliographic value limited to books and material type limited to language materials or to e-books. We created a file for each fiscal year.

Data Collection

Fail: Order Records

Win: Item Records
Created

Line	Type	Field
1	b	OCLC #
2	BIBLIOGRAPHIC	RECORD #
3	BIBLIOGRAPHIC	TITLE
4	BIBLIOGRAPHIC	STANDARD #
5	BIBLIOGRAPHIC	IMPRINT
6	BIBLIOGRAPHIC	CALL #
7	BIBLIOGRAPHIC	LANG
8	BIBLIOGRAPHIC	LOCATION
9	BIBLIOGRAPHIC	MATTYPE
10	BIBLIOGRAPHIC	BIBLVL
11	BIBLIOGRAPHIC	NOTE
12	BIBLIOGRAPHIC	SUBJECT
13	ORDER	ORD TYPE
14	ORDER	VENDOR
15	ORDER	PAID
16		

Paid Field will be limited on Paid Date
Date range: 07-01-2014 to 06-30-2015

CLAREMONT COLLEGES
LIBRARY • VITeL

Unfortunately, we learned that order records were systematically deleted and re-used until 2012, so we had to begin data collection over again, even after having spent a large amount of time already working on cleaning and normalizing the data.

We started over again with item records created between 1995 - 2015. While we were able to capture much of the information we needed through “item records created,” we had to give up any analysis of vendor, order type, and payment related information.

Data Cleanup & Normalization

1. Removed unusable records
2. Removed diacritics in OCLC numbers
3. Change bib record codes into meaningful words
4. Separate Title and Author
5. Remove government documents
6. Remove non-English titles
7. Save all removed records in a separate file
8. *Upload remaining records to Access database*
9. Query database in the title and subject field for terror*
10. Exported query results into Excel

CLAREMONT COLLEGES
LIBRARY • VITeL

Data Clean-Up & Normalization

1. Removed unusable records (multiple OCLC numbers)
2. Cleaned up OCLC numbers
3. Convert bibliographic record codes into meaningful information
4. Separate Title and Author
5. Separate gov pubs
6. Separate non-English titles
7. Create separate files for gov pubs, non-English titles, and English print books
8. Upload the English print books into an Access database on a shared drive (unsuccessful)
9. Query database in the Subject field for terror* (unsuccessful)
10. Filter Excel files in subject column for Text Contains: terror*) (successful)

Issues: One challenge were the logistics of dealing with large files and uploading into an Access database. We found working offline on a local hard drive, uploading the files into the Access database worked the best – at least for a while. Ultimately, our Access database was unstable because it was placed on a shared drive and we were trying to work with it from shared spaces via Wi-Fi and VPN. We eventually decided to work just from Excel and identify titles on terrorism by filtering each Excel file in the subject column for text containing terror*.

Finalizing a list of item records

- Access database proved unstable
- Decided to use excel
- Searched for terror with a space after
- Ended up with a list of item titles, per fiscal year

CLAREMONT COLLEGES
LIBRARY • VITeL

We downloaded the data for all books purchased between 1995 – 2015 and we took out government docs and took out non-English language titles. Using Excel we filtered for specific terms in both the title field and the subject field of the bibliographic records we collected. Initially we searched for terror*, and later refined to terror_ (terror space) and “terrorism” to capture as much as possible. We had tried to do this by building an Access database and querying on terror* but the database was unstable and difficult manage on a shared drive and we ultimately felt that Excel was more accurate. We hope in the future that we can fully populate the Access database and use this database as a central location for any librarian who wants to query the dataset on subject and title terms for future analysis. Refining our entire data collection of bibliographic records and then refining with an Excel terror* may have resulted in some biases, but these are unknown and warrant further research.

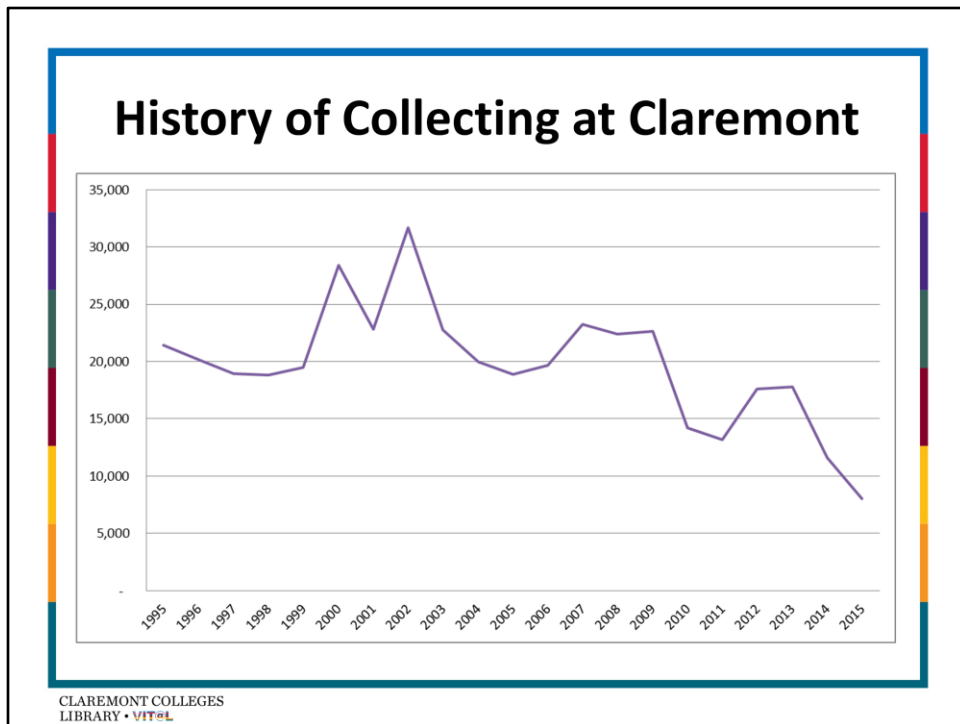
Data Cleanup: The Finger-Mangler



<http://hellinahandbasket.net/wp-content/uploads/2015/01/1901-one-minute-washer-antique-washing-machine-768x1024.jpg>

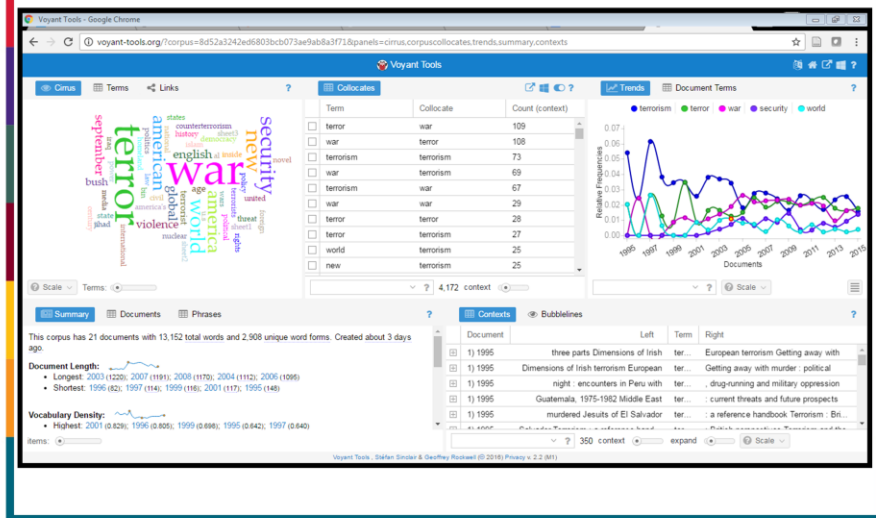
CLAREMONT COLLEGES
LIBRARY • VITeL

It isn't easy to work with new data, especially with new or unfamiliar with, nor collaborate on data-crunching as part of a team. We faced many practical challenges, including crunching large quantities of data on relatively slow machines, and trying to do it on Wi-Fi. We had trouble doing analysis while on VPN. We had random errors in Excel that we weren't able to explain, such as rows that suddenly weren't aligned any more. We had to repeatedly go back and review our data and adjust for errors. In fact, we had to do this over and over, because the closer and more detailed our analysis became, the more we saw anomalies in the data that forced us to go back to the beginning. In a sense, the process looks a lot like the cyclical nature of research itself. It was very frustrating, but also necessary. We would say to anyone who had done data analysis on any scale: don't worry if you run into these sorts of problems because "finger mangling" is part of the process. It is worth it – in fact it is essential – to spend significant time on this phase of any analysis project to ensure your results and conclusions are based on accurate information.



This chart represents the # of item record's created for print books between 1995 – 2015. Although it was disappointing to find out that our library re-used order records, it inspired a series of conversations with long-standing staff members with institutional memory. Through a sort of information oral history, we were able to learn a little bit about the library's collection development history that helped us to put our collecting trend lined into context. Important information included experimentation on the approval plan between 2010-2012 that explained a drop in book purchasing at that time. We also migrated to a new ILS in July of 2015 (we moved from Millennium to OCLC's WMS). In the months leading up to the migration, print ordering was temporarily stopped. At the same time, we started a new approval plan the same summer.

Analysis with Voyant Tools



CLAREMONT COLLEGES
LIBRARY • VITeL

So after coming to grips with the data we are working with, we decided to visualize the content of the titles.

Voyant tools is a web based text reading and analysis environment. It is designed for you to see through your text for patterns.

We created separate files for each fiscal year, and then loaded that collection of 21 files into voyant.

Voyant then uses a number of tools, called skins, where you can investigate different elements of your textual data.

We chose five – Cirrus, Corpus Collocates, Trends, Summary, and Contexts. Cirrus creates clouds based on frequency, summary gives you numbers and lists of distinctive words, and contexts gives you a number of words around each word.

Here is our dashboard: <http://voyant-tools.org/?corpus=8d52a3242ed6803bcb073ae9ab8a3f71&panels=cirrus,corpuscollocates,trends,summary,contexts>

Experimenting with Collocation

Voyant Tools Collocates

Term	Collocate	Count (context)
Islam	This is the keyword term around which collocate (context) terms are counted.	8
Islam	war	5
Islam	new	5
Islam	terror	4
Islam	global	3
Islam	world	2
Islam	violence	2
Islam	understanding	2
Islam	swift	2
Islam	struggle	2
Islam	rage	2
Islam	protecting	2
Islam	press	2
Islam	myths	2
Islam	muslim	2
Islam	ihad	2
Islam	islamiyah	2

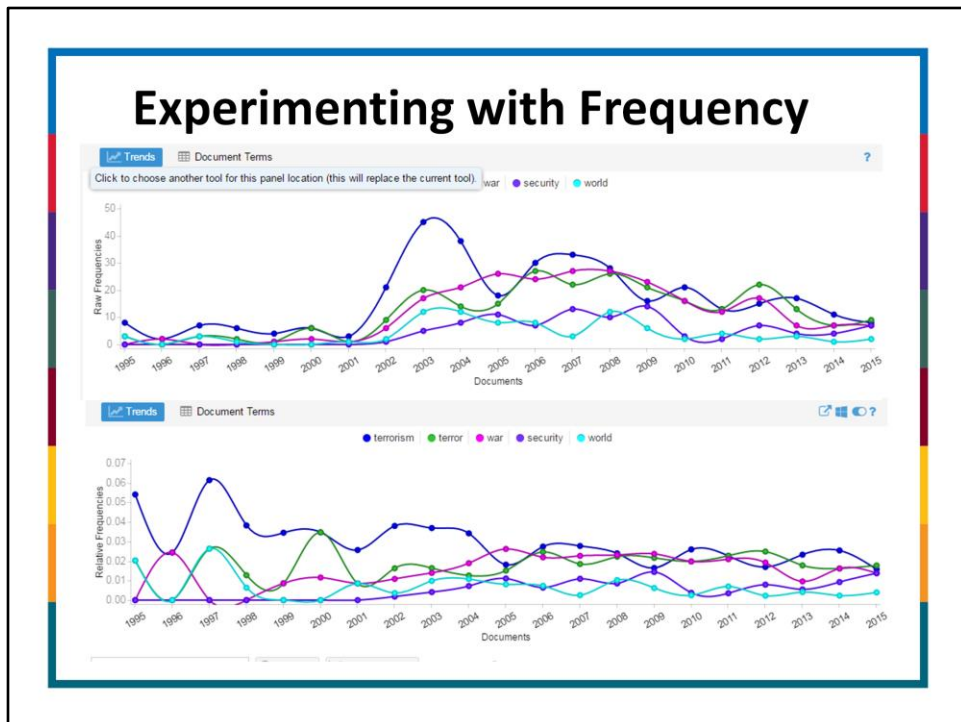
Voyant Tools Collocates

Term	Collocate	Count (context)
christianity	trialogue	1
christianity	terror	1
christianity	islam	1
christianity	battlefield	1

CLAREMONT COLLEGES
LIBRARY • VITeL

We used Corpus Collocates to get a sense of connections between different words – it will sort the corpus into terms, collocated (nearby) terms, and the frequency of when those terms are together. You can select a certain term – such as “Islam” or “Christianity” and see what words are nearby to get a snapshot of what concepts are being used near each other.

If you want to stabilize this in a visualization, you can export as tab separated values and copy and paste into an excel sheet. Clean up excel file and add headers of “Term” “Collocate” and “Count”.



Another tool used was the trend lines tool, which represents frequencies in terms across segments in a document, which here is per fiscal year. The tool automatically pulls out the top words, and maps them out.

Voyant allows you to look at both raw frequency, which is the number of times a title occurred in any given FY (top), or relative frequency, which is term frequency in a segment per normalized count of one million items.

Useful for getting a snapshot of how terms in titles occur over time – not just item frequencies but getting an idea of context.

Drawbacks of Voyant Tools



A note of warning:

Web based, so sometimes slow and possibly unstable. Might not load, even if you have an embed code.

Difficulty with exporting – best for visually analyzing your data within voyant, rather than exporting it

Because of this, we ended up exporting the data out of Voyant for a few other visualizations.

[illegible]

Aside from taking raw frequencies from Voyant, we used two other tools.

We used excel filtering to isolate most frequent terms in the datasets.

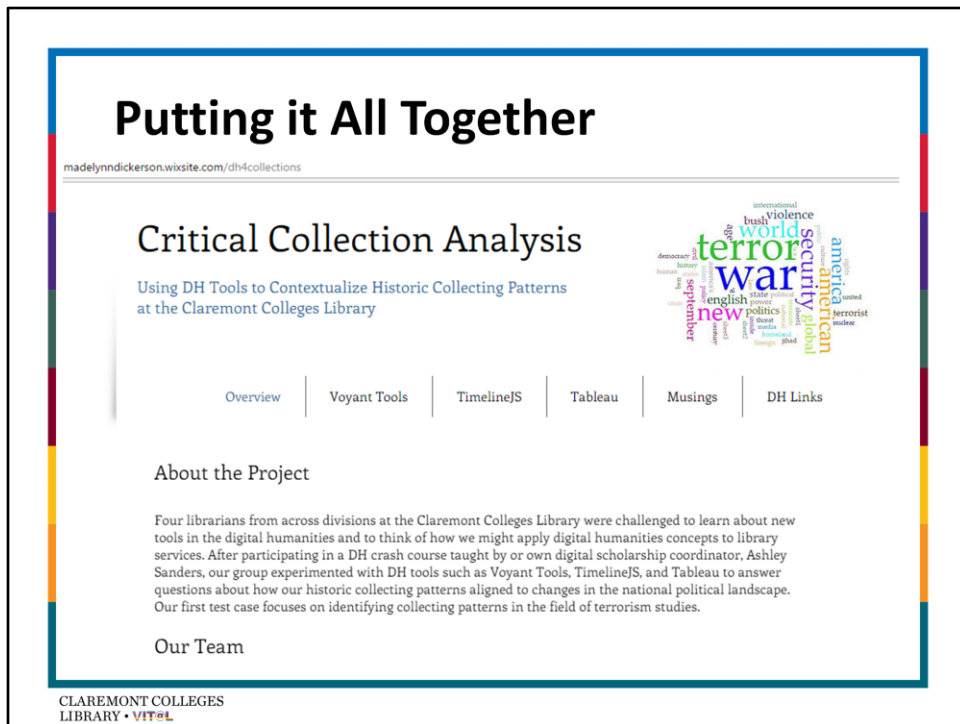
We used Python, a programming language, to analyze frequent terms using a different tool. We decided on this tool because we wanted numbers on frequencies of different tools, but Voyant would not export the numbers we needed.

With the (major) assistance from our Digital Scholarship Coordinator Ashley Sanders we ran a word frequency analysis on the excel files that contained book titles, saved the results as a text file, and then converted that to excel files in order to have data on what words showed up in our titles and their frequency. This may seem like overkill, but it could serve to be potentially very useful for larger data sets.

Fiscal Year	Excel filtering Most Frequent Title Term	Voyant Most Frequent Title Term	Python Most Frequent Title Term
1995	terror	terrorism	terrorism
1996	terror	world	war, terrorism, oklahoma, city (tie)
1997	terror	terrorism	terrorism
1998	terror	terrorism	terrorism
1999	terror	terrorism	terrorism
2000	terror	terror	terrorism
2001	terror	terrorism	terrorism
2002	terror	terrorism	terrorism
2003	terror	terrorism	stories, other, new, history, dying (tie)
2004	terror	terrorism	terrorism
2005	terror	war	war
2006	terror	terrorism	terrorism
2007	terror	terrorism	terrorism
2008	terror	terrorism	war
2009	terror	war	war
2010	terror	terrorism	terrorism
2011	terror	war	terrorism
2012	terror	terror	terror
2013	terror	terrorism	terrorism
2014	terror	terrorism	terrorism
2015	terror	terror	terror

We used three different tools to look the most frequent terms used: Voyant, excel, and Python.

Using Excel filtering, Terror was the most frequent term in title and subject except for 2012, when war was the most frequent term in title. But when we ran a python script we got different words. This is a subject that needs further investigation.



We ultimately decided to put everything together in an online portal that we created using Wix.

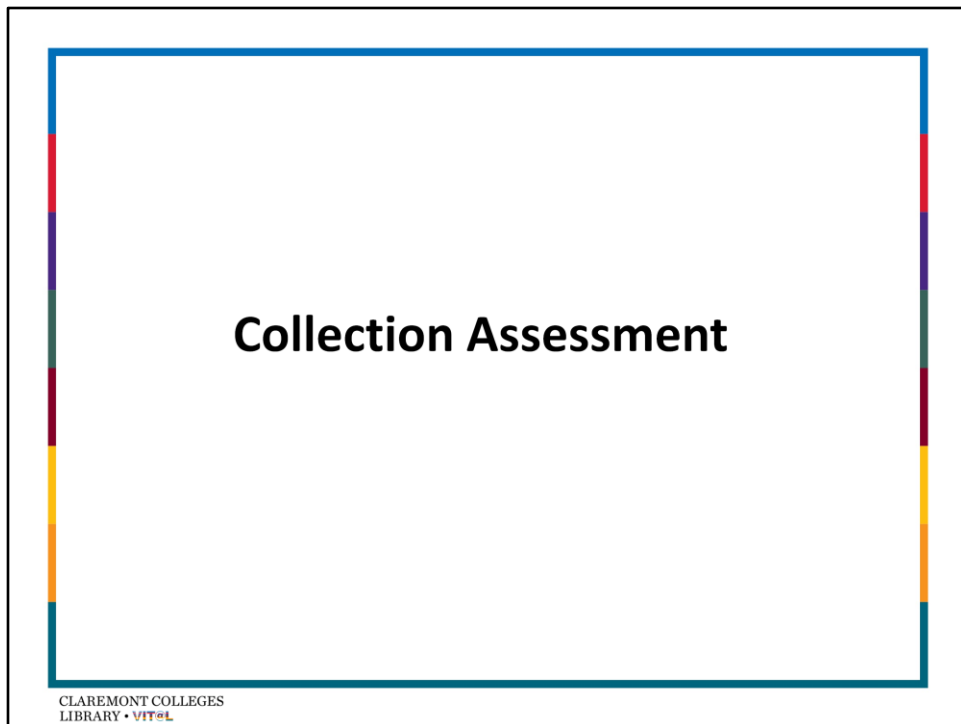
We investigated Timeline JS as a way to incorporate all our data for visual analysis.

Created list of terror events using shared knowledge and chronology from the [National Counterterrorism Center](https://www.fbi.gov/ncct). Created timelines for collection development events at Claremont Colleges Library, new print titles, and distinctive terms from voyant.

Timeline JS was originally developed for journalism purposes, to develop a story that is visually rich and interactive. Although we were able to load all our information in TimelineJS, it did not add much value to our analysis and actually hid our story.

Link to our project:

<http://madelynndickerson.wixsite.com/dh4collections> OR bit.ly/claremontcollections



After exploring text analysis with Voyant and data/information visualization with TimelineJS and Tableau, we definitely see the potential for DH tools in the future of collection assessment especially when used in conjunction with traditional tools and methods. Many collection assessment projects ask questions about “how many” titles and in what subject areas. Or “how much” titles cost and “how much” they were used. These are important quantitative assessments. What text-analysis provides is an opportunity to capture the nuances of a collection, especially interdisciplinary collections that cross many LC call numbers. We also see a growing emphasis on interactive data visualizations that allow librarians across divisions to explore collections in context. A dashboard similar to what we have created on the website could serve as a useful means of introducing a collection to a new faculty member or a new subject librarian. Thinking about historical trends in collection building at an institution, and communicating these trends through timelines, can inform future collection development approaches in a meaningful way. There is a great deal more work to be done to refine DH approaches to collection assessment, but we believe it is worth the investment in time and effort to do so.



Try methodology with new terms, especially with terms that have changed over time.
Introduce collections to new faculty.

Good tool for partnership with scholars in different topics and we can build relationships – DH is about partnerships

A good way to introduce liaison librarians to new subject areas or to familiarize librarians with collections.

In Tableau we can potentially add title lists in the tool tip to create an alternative form of collection browsing.

DH is about experimentation and we had the luxury to spend time exploring and following the natural flow of our ideas. This isn't necessarily something everyone has, but we hope that our project can eventually serve as a stepping stone for others.

Things to think about

- Data “hygiene”
- Tool choice
- Security
- Cost
- Storage

Data management – be careful with your data. Keep a watchful eye all the way throughout your process. For example our columns shifted. Think about naming conventions and file management, especially when sharing data or working collaboratively, which adds an extra layer of challenge.

Tool Choice – decide what you need to do, and pick your tool appropriately rather than trying to make your project fit a tool. It doesn’t have to be fancy. If you want to know your top publishers, Excel might work the best!

Security – Voyant stores tools to make them available in different work sessions. Can your data be stored publically?

Cost – the tools we used in this project are free and available for public use. In exchange our data is freely viewable, and stored on a server that we do not necessarily have control over. These are decisions you need to make if you’re doing something similar.

Storage – where are you planning on storing your project?

Our Website Link

bit.ly/claremontcollections

CLAREMONT COLLEGES
LIBRARY • **VITeL**

The slide features a white background with a blue border. The title 'Our Website Link' is in bold black text at the top. The URL 'bit.ly/claremontcollections' is centered in black text. The bottom left corner contains the text 'CLAREMONT COLLEGES LIBRARY • VITeL' in a smaller font. The border is decorated with vertical bars of red, purple, green, yellow, and orange on the left and right sides.

As this website was a work in progress designed to document our analysis as it was on November 4, 2016 – the website may have changed since this presentation was originally given.

Contact Us

Lydia Bello

STEM Team Librarian

lydia_bello@cuc.claremont.edu

Madelynn Dickerson

Information Resources Coordinator

madelynn_dickerson@cuc.claremont.edu

Nina Clements

Social Sciences Team Librarian

nina_clements@cuc.claremont.edu

Margaret Hogarth

Information Resources Acquisitions

Team Leader

margaret_hogarth@cuc.claremont.edu

CLAREMONT COLLEGES
LIBRARY • **VITeL**